Segmental Influences on the Perception of Pitch Accent Scaling in English

Jonathan Barnes[1], *Alejna Brugos*[1], *Nanette Veilleux*[2], *Stefanie Shattuck-Hufnagel*[3]

[1] Boston University, Boston, Massachusetts, USA
[2] Simmons College, Boston, Massachusetts, USA
[3] Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
jabarnes@bu.edu, abrugos@bu.edu, veilleux@simmons.edu, sshuf@mit.edu

Abstract

In both tone and intonation systems, segmental context is known to influence production and perception of target F0 contours in various ways. Many languages, for example, prefer to realize critical F0 events during maximally sonorous intervals, either by varying the timing of pitch movements, or by virtue of distributional limitations on certain contour types. Current analytic practice, by contrast, routinely ignores segmental backdrop when estimating the perceptual efficacy of putative cues, such as F0 turning points, to tone scaling and timing patterns. Results of the perception study presented here argue that pitch accent scaling is best modeled using a weighted average of F0 sampled over a defined region of interest, and that individual sample weights are determined in part by the sonority of the segments from which they are taken. That is, samples from lower sonority segments contribute less to integrated scaling percepts than those from higher sonority segments. This model, called TCoG-F(requency), accounts for crosslinguistic tonal timing and distribution patterns in the literature, and underscores the danger of analyzing tonal phenomena completely apart from the segments that express them.

Index Terms: Intonation, Pitch perception, tone scaling, tonal timing, sonority, Tonal Center of Gravity.

1. Introduction

Since the dawn of the autosegmental era in tonal and intonational phonology [10, 19, 4, 22], we have grown accustomed to thinking of linguistic pitch specifications as existing apart from, or parallel to, the segmental skeleton of the spoken utterance. Specifications on the so-called tonal tier must then be associated, according to the dictates of the grammar, with appropriate Tone-Bearing Units in order to be realized phonetically. At the same time, however, it is well known that both perception and production of F0 can be influenced significantly by the segmental contexts in which contours are realized. For example, 'microprosodic' effects on the F0 contour stem from, e.g., voicing differences in syllable-initial consonants, differences in vowel height between otherwise comparable syllables, etc., and experimentalists routinely control for such effects [15, 17, 26].

Less widely appreciated, however, are the ramifications of a commonly-remarked tendency for languages to avoid realizing critical portions of F0 contours within lower-sonority regions of the segmental string. This tendency manifests itself in various ways: On the one hand, intonational phonologists have observed what appear to be systematic alterations to tonal timing patterns in order to ensure optimal expression of F0 contours in a given segmental context. For example, in a variety of languages, accentual High F0 targets occur relatively earlier in closed syllables than in open, and in syllables closed by obstruents than in those closed by sonorants [5, 18, 23, 24, 25]. On the other hand, languages can impose categorical distributional restrictions on the association of certain tone patterns with particular kinds of segment hosts. For example, cross-linguistically, contour tones tend to be restricted to syllables with longer, higher-sonority rhymes [11, 28, 29].

The latter pattern, typically observed in languages with lexical tone contrasts, has been explained as resulting from the comparatively greater salience of the percept of pitch during segments that are higher in intensity and richer in harmonic structure [8, 29]. Under this scenario, languages deploy their fullest array of tonal contrasts only in contexts where these contrasts are most likely to be accurately perceived. This explanation could account for the first cases mentioned above as well: If F0 peaks associated with intonational High pitch accents occur relatively earlier in closed syllables than in open, for example, we might attribute this to speakers' desire to realize critical pitch information (i.e. the bulk of elevated F0) within a more sonorous portion of the syllable (i.e. the nucleus rather than the coda).[1]

Against this backdrop of general awareness of the influence of segmental context on tone perception and production, there is also a somewhat paradoxical countervailing tendency to assume that putative tonal targets such as F0 turning points are of equal perceptual value regardless of the nature of their host segments. According to this practice, linguistically meaningful pitch accent scaling patterns are equated phonetically with measured F0 maxima, regardless of where in the segmental string those maxima fall; similarly, the temporal location of F0 turning points is estimated according to the visual salience of 'corners' in the F0 track, regardless of whether those corners fall in regions of the signal with high or low auditory salience. In an analogous vein, the subjective continuity of intonation contours, even through regions where F0 is heavily disrupted by intervals of voicelessness, has led some to assume that listeners have an ability to effectively 'restore' missing F0 intervals to the signal via interpolation, or extrapolation based on existing trajectories [12, 13, 21]. Accordingly, F0 stylization algorithms such as the Fujisaki model, MOMEL, or Tilt [9, 14, 27] create continuous F0 tracks based on gappy originals, in some cases even locating critical F0 target points within such 'filled in' intervals, when the shape of the interpolated pitch curve suggests it. It is the tension between this tendency, on the one hand, and the literature on avoidance of low-

[1] Another explanation that has been offered for these altered timing patterns is based on House's Spectral Stability Hypothesis [16, 7, 24], whereby pitch movements are more readily perceived as such when they are realized during regions of spectral stability. For reasons of space, we will not treat this hypothesis further here, other than to note that it relates less obviously to the distributional restrictions on lexical tones noted above.

10.21437/SpeechProsody.2014-214

sonority segmental hosts on the other, that inspired the experiments described in this paper.

A first step toward resolving these issues was taken recently by [1], who demonstrate that, at least for perceived F0 target scaling, the "perceptual completion" approach cannot be correct.[2] In that study, subjects made judgments regarding the scaling of synthetic English High pitch accents (L+H*) realized either as clear peak- and plateau-shaped F0 contours extending over fully voiced segmental intervals (e.g., in a context like '*DAY' might fit*), or as analogous contours in which the region corresponding to the nuclear pitch accent contained 'missing' or inferable peaks/plateaux (i.e. mirror-image rises and falls separated by the closures and releases of voiceless stops, as in a context like *'DATE' might fit*). Rather than either extrapolating or interpolating F0 across such voiceless intervals in a way that would register systematically on scaling judgments, subjects were seen to behave as though the missing intervals were absent altogether, ignoring the gaps, and judging relative pitch accent scaling exclusively on the basis of the F0 values actually present in the signal.

While that study gives an indication of how listeners treat F0 gaps created by voiceless stops, it remains unclear what listeners do with intervals in which F0 is in fact present, but with lower-amplitude or spectral impoverishment. In the current study we use similar investigative techniques to those of (1), applied specifically to the perception of measurable F0 over lower-sonority intervals. We hypothesize that vowel vs. silence are in fact two ends of a continuum of possible F0 carriers with differing degrees of salience, along which higher sonority segments such as liquids or nasals give way gradually to lower sonority ones, such as voiced fricatives or stops. We predict that these sonority-based differences in the robustness of perceived pitch will be manifested in listeners' scaling judgments. We situate the results of this study in a model of tone scaling perception we call TCoG-F (Tonal Center of Gravity in the Frequency dimension), where the perceived scaling of an F0 event (e.g., the elevated F0 associated with a High pitch accent) is modeled as a weighted average of F0 measured over a particular region of interest. In the calculation of this average, F0 samples taken from more sonorous regions are accorded heavier weights, and are thus predicted to extend relatively greater influence over perceived scaling.

2. Methods

The reasoning behind the current design is as follows: Consider a set of utterances, such as those depicted in Figure 1, bearing rise-fall-rise intonation contours, with plateau-shaped L+H* nuclear pitch accents on the first word of the sentence (*'X' might fit*, uttered perhaps in the context of the solution to a crossword puzzle clue), where F0 rises through the nuclear vowel, remains high and level through the coda nasal, and falls thereafter, before rising at the end to signal something like tentativeness. In all three examples, the syllable rhymes (and their constituent nuclei and codas in a. and c.) are identical in duration, as are the relevant segments of F0 contour. What differs is only the sonority of the portion of the syllable rhyme bearing the high, level portion of the accentual plateau for the three target words *Dane*, *day* and *Dave*. Assuming, as hypothesized above, that the perceived scaling of this pitch accent involves averaging over F0 samples taken during the entire rhyme of the accented syllable, we expect first that the perceived scaling of all three of these words will end up lower than the maximum F0 realized during the plateau-portion of the contour. However, to the extent that the perceptual contribution of any given F0 sample is weighted by a factor representing the sonority of the segment bearing it, we expect the perceived scaling of the pitch accents in these three utterances to differ from one another as well. Since the highest portion of the pitch accent in *day* (i.e. the plateau) occurs in a region of greater sonority than the analogous portion of the pitch accent in *Dane*, the high F0 samples for *day* will contribute more to the resulting average, so that *day* will sound higher to listeners than *Dane,* despite identical F0. Correspondingly, owing to the lower sonority of this region in *Dave,* we predict the pitch accent in 1c to sound lower than those in a and b, again despite the lack of difference in objective F0.

To test these predictions, we designed a set of experimental stimuli similar to those used by [1] but differing in several critical ways. As just described, all stimuli were instances of the English words *day, Dane,* and *Dave,* realized in the target position of the frame sentence *X might fit*. All target stimuli were realized with rise-fall-rise intonation contours (ToBI L+H* L-H%), with the nuclear pitch accent on the first word. However, F0 for these base utterances was resynthesized to create contours of two basic types: plateau-shaped pitch accents and sharp-peak-shaped pitch accents, in effect extrapolating the preceding rise and following fall to a single higher intersection point instead of a plateau. These are both depicted in Figure 1, and their deployment in our task is detailed in Section 2.2 below.

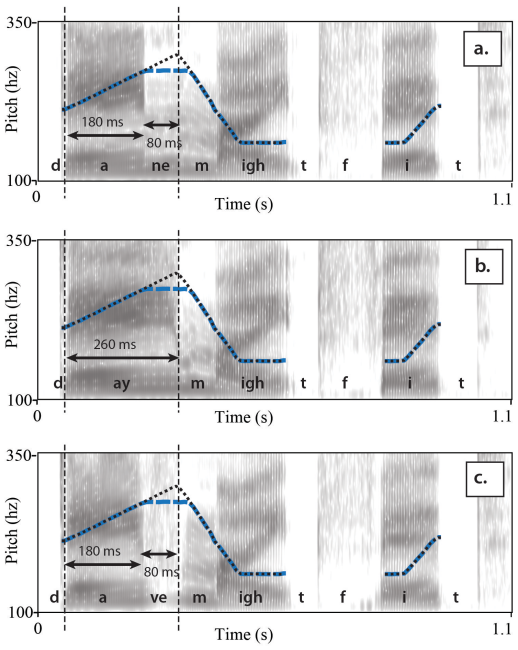

Figure 1: *Spectrograms with superimposed pitch tracks for plateaux (blue dashed lines) and peaks (black dotted lines) for* Dane *(a),* day *(b) and* Dave *(c).*

[2] At least in the most literal sense. It is still possible that non-F0 cues within voiceless regions contribute either to the 'subjective continuity' of the pitch contour, or to the perception of 'prominence' (a linguistic dimension sometimes cued in part by the higher-than/lower-than relations that underly linguistic pitch distinctions [20]. At the same time, is worth noting the extent to which relative pitch and prominence relations vary orthogonally (as in the expression of lexical tone contrasts, or the difference between downstepped and non-downstepped nuclear pitch accents).

2.1. Stimulus Creation

Target phrases were created from two sets of base recordings, one produced by a male native English speaker, the other by a female, and then resynthesized using Praat [3]. Synthesized segment durations for the female speaker, given in Fig. 1, were based on mean values over multiple utterances. For peaks, F0 rises were identical in duration (260 ms) and scaling (212-300 Hz, a 6 st rise) for this speaker for all stimulus types, and were followed by a 140 ms fall to 160 Hz. Plateau stimuli had a rise of the same slope as the peaks, but truncated after 180 ms (212-273 Hz), followed by a 101 ms plateau (at 273 Hz), and a 119 ms fall to 160 Hz (the same slope as in the peak stimuli, but starting from the end of the plateau). Duration and F0 values for stimuli based on the male speaker were comparable, though different in quantitative detail.

2.2. Experimental task

Our primary question pertains to the perceived relative scaling of nuclear L+H* pitch accents realized on syllables with differing rhyme types. However, to avoid the potential for confounds inherent in the direct pairwise comparison of syllables with differing segmental content, the relative scaling of these contours was investigated indirectly. That is, the bulk of experimental trials consisted of the pairing of a given target item (i.e. a target utterance with *day, Dane, or Dave,* with either a peak- or plateau-shaped pitch accent) with one out of a continuum of standard reference contours. These reference contours were segmentally identical to the target item, but F0 throughout the accented syllable was held steady at one of 7 levels, the highest at 300 Hz, descending thereafter in .5 semitone increments (Figure 2). After the fall from the accented syllable, F0 was identical for target items and standards.

If we assume that corresponding level standards sound identical in scaling regardless of syllable type, since sonority-related weighting variation cannot change perceived pitch for a flat-F0 contour, then any effects of F0 weighting differences on tokens with changing F0 should be manifest in subjects' perception of the relative scaling of target items and their respective level standards. (E.g., *day* might sound equal in pitch to its standard level 5, while *Dave* might reach only level 4.)

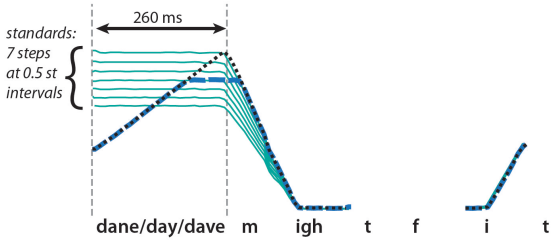

Figure 2: *Pitch tracks for standards (teal solid), plateaux (blue dashed) and peaks (black dotted).*

The task itself was 2AFC: 77 native speakers of American English were presented with pairs of contours (a target item and a level standard), and decided which contour's target word reached a higher pitch. After 6 consecutive correct responses in an initial block comparing standards separated by ≥ 3 steps, the experiment began, with each test item (2 accent shapes X 3 word types) compared to its continuum of 7 standards (2 reps of standards 1, 2 & 7, and 3 reps of standards 3, 4, 5, & 6 = 18) in 2 orders, for a total of 216 trials. Additionally, there were two more trial types interspersed. The first was represented by 36 trials pairing two level standards separated by either 2 or 3 continuum steps for *day* target types (18 comparisons x 2 orders). These trials served as a baseline measure of participants' accuracy in discriminating pitch levels. The final trial type involved 18 pairings consisting of each sharp-peak-shaped version of a given syllable type with its plateau-shaped counterpart (3 reps X 2 orders X 3 word types). These trials served as an additional test of the hypothesis that the phenomena under investigation here are in fact the result of lowered perceptual salience of F0 samples taken from utterance intervals of lesser sonority. In all such pairings, as with the level vs. level comparisons, there is in fact a "correct" answer: The sharp peak version of the pitch accent should sound higher than its plateau-shaped counterpart because it is, in fact, by any measure, higher.[3] On the other hand, since the entirety of the region of F0 difference between peaks and plateaux fell within the region of sonority difference across word types, if our hypothesis is correct, the scaling difference between the two should be relatively easy to detect for *day,* but progressively harder in the lower sonority rhymes of *Dane-* and *Dave*-type stimuli. All trial types were mixed together, and all 270 experimental trials were presented in random order, with breaks after every 50 trials. (See Figure 3 for a schematic summary of the 4 trial types.)

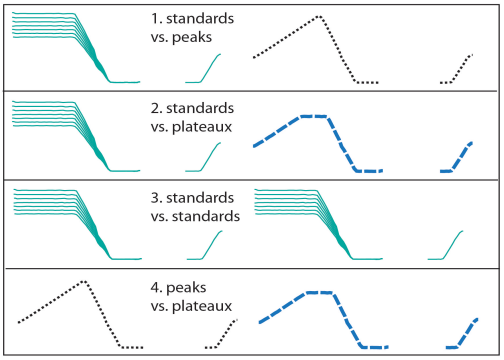

Figure 3: *Schematic showing the 4 types of experimental trials.*

2.3. Results and analysis

Data from 62 participants is included in the analysis. (Of 77 total, 15 of whom did not reach criterion for inclusion based on discrimination of level standards.) Fig. 4 displays results, pooled across subjects. Lines represent the percentage of trials in which *day, Dane,* or *Dave* was judged higher than each of its 7 level standards. Comparing target types, the percentage of 'higher-than' judgments for *Dave* clearly declines earlier in the continuum of level standards than does that of *Dane,* which in turn declines earlier than *day*. We infer from this that *day* sounds higher to listeners than *Dane,* and that *Dave* sounds lower. This is confirmed by a mixed-effects logistic regression analysis, using both standard level and target-syllable type, as well as accent shape (peak vs. plateau) as fixed factors, and participant as a random factor. The resulting model (N = 12,971, log-likelihood = -5573) shows a main effect of standard level (Est. = -0.934 (SE = 0.017), Wald Z = -55.73, p

[3] Mean F0 for the accentual high region, however measured and weighted, was higher for peaks than for plateaux.

< .001), and of word type, with *day* differing in a positive direction from *Dane* (Est. = 0.256, (*SE* = 0.079), Wald *Z* = 3.24, *p* = .001), and *Dave* differing from *Dane* in a negative direction (Est. = -0.504 (*SE* = 0.08), Wald *Z* = -6.29, *p* < .001). Importantly, in addition to a main effect of accent shape, with plateaux differing in a negative direction from peaks (Est. = -0.941 (*SE* = 0.082), Wald *Z* = -11.35, *p* < .001), there was also a significant interaction between accent shape and word type: the peak-plateau difference was less salient for *Dave* than for *Dane* (Est. = 0.276 (*SE* = 0.118), Wald *Z* = 2.34, *p* < .05) (*day* did not differ significantly from *Dane* in this respect).

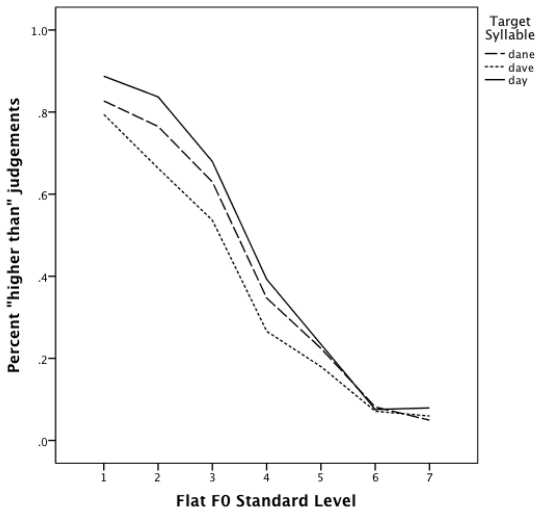

Figure 4: *Percent 'Higher-than' judgments for the three syllable types, as a function of the level standard against which they were compared.*

These results strongly bear out the predictions of the TCoG-F hypothesis detailed above. *Day* tokens sounded higher than *Dane,* which sounded higher than *Dave.* Critical in understanding this primary result is the significant interaction between accent shape and word type. The fact that the perceived scaling difference between peaks and plateaux was less pronounced when realized over the voiced fricative in *Dave* than in the nasal coda of *Dane* suggests that pitch percepts stemming from the former region are indeed less robust than those originating in the latter.

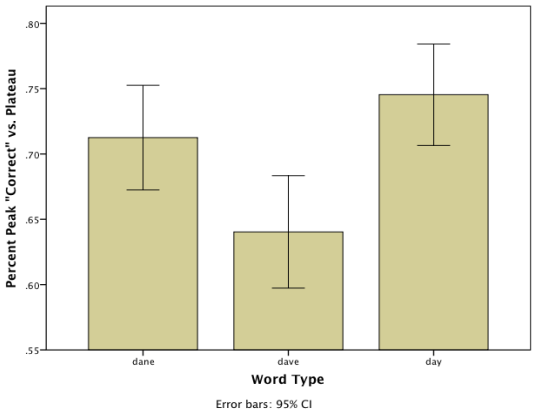

Figure 5: *Percent of peak vs. plateau trials in which peaks were correctly judged to sound higher than plateaux, for three target syllable types.*

This same conclusion is supported by trials in which subjects compared peak-shaped versions of a contour directly with their plateau analogues (Figure 5). Listener judgments of these comparisons were most accurate (i.e. listeners heard peaks as higher) for *day,* less so for *Dane*, and were least accurate for *Dave.* Another mixed-effects logistic regression (with word type as a fixed effect and participant as a random effect, *N* = 1462, log-likelihood = -849.4) shows that, while the *day-Dane* distinction was non-significant (Est. = 0.187 (*SE* = 0.15), Wald *Z* = 1.248, *p* = .21), the distinction between *Dave* and *Dane* was significant in the predicted direction (Est. = -0.358 (*SE* = 0.144), Wald *Z* = -2.494, *p* = .012). This lowered accuracy for scaling judgments where the sole difference between the two F0 contours lies within the lower sonority region of the coda is, again, just what TCoG-F would predict: Lower F0 sample weights during the crucial interval understate the objective difference between the two contours and make judgments more error-prone.

It is also worth noting that the connection between this result and the earlier one is a first glance not obvious: on the face of it, the fact that *Dave* trials sounded systematically lower than *Dane* or *day* to our listeners when paired with level standards bears no logical connection to the degree of accuracy listeners might exhibit when comparing one kind of *Dave, Dane,* or *day* contour to another in paired scaling judgments. The connection between the two results becomes clear only through the lens of TCoG-F.

3. Conclusions

While from a phonological point of view, the advent of autosegmentalism brought with it a great deal of progress and innovation, in the domain of phonetic realization, there remains a persistent danger that the core autosegmental insight, the separation of tonal and segmental phenomena onto distinct representational "tiers", may at times be taken too literally. The findings described here suggest that listeners' perception of F0 in speech signals is influenced by the nature of the consonant and vowel segments over which the F0 pattern occurs, even when voicing continues through those segments; that is, F0 values in regions controlled by more-sonorous segments (like vowels and nasals) are weighted more heavily than F0 values in regions controlled by less-sonorous segments (like voiced fricatives). This means that models based on a straightforward mapping between values of F0 peaks and valleys on the one hand, and perceived intonational targets, on the other, will need to be modified to take account of the influence of host segments. Taken together with earlier findings demonstrating the role of F0 contour shape on perceived tonal target alignment [2], these results support a model of intonation processing based on the Tonal Center of Gravity in both the time and frequency domains. We suggest furthermore that perceptual registration of individual cues in the speech signal is only the first step in the process of integrating multiple cues to form a single linguistically-relevant auditory percept. Future work will test this hypothesis in languages other than American English.

4. Acknowledgements

We gratefully acknowledge the support of NSF grants 1023853, 1023954, and 1023596.

5. References

[1] Barnes, J., A. Brugos, N. Veilleux & S. Shattuck-Hufnagel. 2011. Voiceless intervals and perceptual completion in Fo contours: Evidence from scaling perception in American English. *Proceedings of the 17th International Congress of Phonetic Sciences, August 2011, Hong Kong.*

[2] Barnes, J., N. Veilleux, A. Brugos, & S. Shattuck-Hufnagel. 2012. Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Journal of Laboratory Phonology* 3(2): 343-389.

[3] Boersma, P. & D. Weenink. *Praat: doing phonetics by computer*, accessed May 1, 2009. http://www.praat.org.

[4] Bruce, Gösta. 1977. *Swedish word accents in a sentence perspective*. (Travaux de l'Institut de Linguistique de Lund 12.) Lund, Sweden: CWK Gleerup.

[5] Caspers, J. & V. van Heuven. 1993. Effects of time pressure on the phonetic realization of the Dutch accent-lending pitch rise and fall. *Phonetica* 50: 161-171.

[6] d'Alessandro, C. & P. Mertens. 1995. Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language* 9: 257-288.

[7] Dogil, G & A. Schweitzer. 2011. Quantal effects in the Temporal Alignment of Prosodic Events. *Proceedings of the 17th International Congress of Phonetic Sciences, August 2011, Hong Kong.*

[8] Flemming, E. 2008. The grammar of coarticulation. To appear in M. Embarki & C. Dodane (eds.), *La Coarticulation: Indices, Direction et Representation*.

[9] Fujisaki, H. & K. Hirose. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan (E)* **5**(4): 233–241.

[10] Goldsmith, J. 1976. Autosegmental phonology. PhD Thesis, MIT.

[11] Gordon, M. 1999. Syllable weight: Phonetics, phonology, typology. PhD Thesis, UCLA.

[12] Hermes, D. 1998. Auditory and visual similarity of pitch contours. *Journal of Speech, Language, and Hearing Research* 41(1): 63-72.

[13] Hermes, D. 2006. Stylization of Pitch Contours. In S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, & J. Schließer (eds.), *Methods in Empirical Prosody Research,* Berlin-New York: de Gruyter, 29-62.

[14] Hirst, D. & R. Espesser. 1993. Automatic modelling of fundamental frequency using a quadratic spline function. Travaux de l'Institut de Phonétique d'Aix 15: 71–85, Univ. de Provence.

[15] Hombert, J. M., J. Ohala & W. Ewan. 1979. Phonetic explanations for the development of tones. *Language* 55: 37-58.

[16] House, D. 1990. *Tonal perception in speech*. Lund, Sweden: Lund University Press.

[17] Kohler, K. 1990. Macro and micro F0 in the synthesis of intonation. In J. Kingston & M. Beckman (eds), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech,* Cambridge: CUP, 115-138.

[18] Ladd, D. R., I. Mennen & A. Schepman. 2000. Phonological conditioning of peak alignment in rising pitch accents in Dutch. *Journal of the Acoustical Society of America* 107: 2685-2696.

[19] Leben, W. 1973. Suprasegmental phonology. PhD Thesis, MIT.

[20] Mixdorff, H & O. Niebuhr. 2013. The Influence of F0 Contour Continuity on Prominence Perception. *Proceedings of Interspeech 2013, Lyon, France.*

[21] Nooteboom, S. 1997. The prosody of speech: Melody and rhythm. In W. Hardcastle & J. Laver (eds.), *The Handbook of Phonetic Science,* Oxford: Blackwell, 640-673.

[22] Pierrehumbert, J. 1980. The Phonetics and Phonology of English Intonation. PhD Thesis, MIT.

[23] Prieto, P. 2009. Tonal alignment patterns in Catalan nuclear falls. *Lingua* 119 (6): 865-880.

[24] Prieto, P. & F. Torreira. 2007. The segmental anchoring hypothesis revisited. Syllable structure and speech rate effects on peak timing in Spanish. *Journal of Phonetics* 35(4): 473-500.

[25] Santen, J. van & J. Hirschberg. 1994. Segmental effects on timing and height of pitch contours. *Proc. ICSLP 94*, Yokohama, 719-722.

[26] Silverman, K. 1987. The structure and processing of fundamental frequency contours. PhD Thesis, University of Cambridge.

[27] Taylor, P. 1998. The Tilt intonation model. In R. Mannell & J. Robert-Ribes, (eds.), *Proc. ICSLP 98*, volume 4, 1383–1386.

[28] Zhang, J. 2001. The Effects of Duration and Sonority on Contour Tone Distribution— Typological Survey and Formal Analysis. PhD Thesis, UCLA.

[29] Zhang, J. 2004. The role of contrast-specific and language-specific phonetics in contour tone distribution. In B. Hayes, R. Kirchner & D. Steriade (eds.), *Phonetically-Based Phonology*, Cambridge: Cambridge University Press, 157-190.